THE MEASUREMENT OF THE EFFICIENCY OF MENTAL TESTS[1]

BY BEARDSLEY RUML

University of Chicago

The value of a mental test in selecting from a group of individuals those possessing certain academic or vocational abilities is indicated by the amount of relationship which has previously been found to exist between the standing of members of a similar group as given by performances in the test, and their standing in the ability tested for, as determined by the estimation of persons who are qualified to judge, or from the actual achievements of the individuals. In the academic field, when tests are given for the purpose of discovering the mental abilities of students, the standings of the students on the basis of the tests must be related with grades, or with the judgments of ability as given by teachers, before the working value of the tests can be stated. In business and industry, when tests are given to applicants for employment in order that those who are best fitted may be chosen, the efficiency of the tests can be judged only after the results as given by the tests have been compared with the abilities of the candidates, as shown by their subsequent achievements or by the judgments of their superiors.

This relationship, by which the value of a mental test is known, is too complex to be grasped by mere inspection of the data, and symbols of various kinds have been devised to aid in the interpretation of the facts. The most satisfactory measure of the relationship is the coefficient of correlation. As a result, the coefficient of correlation has been widely used in psychology as an indication of the worth of a test; yet when it is computed by the more common methods, *i. e.*, the product-moment method, the method of rank differences,

[1] Mental tests for 'intelligence' and for 'general ability' are to be distinguished from tests in the school subjects, such as algebra and reading tests.

and the foot-rule, it gives an erroneous idea of the true value of the test *in a certain kind of practical situation.*

In these situations to which we refer, mental tests have their greatest use as means of preliminary classification of individuals. The practical problem is to separate individuals into roughly homogeneous groups which will wait for their final internal arrangement upon the development of unmeasurable personal qualities. If a test is to be used in such situations, its value must be determined with reference to these situations. To correlate performance in a test with the exact evaluation of each individual's ability (the method of the three commoner formulæ) is to measure the test in terms of a problem which the test will never be called upon to solve—namely, the determination of the precise ability of each individual. If in the classroom, for example, all that is desired is the separation of students into fairly distinct classes of *good*, *mediocre*, and *poor*, there is no demand upon the test to rank the students in their actual order of merit. The final order of the students will be determined partly by factors which are clearly non-intellectual, and which no test would be expected to anticipate. In business the same condition obtains. The purpose of the test is fulfilled if it succeeds in merely picking out the applicants who are superior, allowing their industry and moral qualities to fix their final rank.

A still more concrete case may be worth describing. Suppose that we have ranked 500 college freshmen, first according to estimates of their ability by their instructors, and second according to their performances in a series of mental tests. The correlation between these standings, let us say, is +.50. The following year we might want to pick out freshmen at the beginning of the school year for advanced divisions in the freshman subjects. The question arises, "Is it possible to make the selection of the brighter students on the basis of their performances in the tests?" Although we know that the correlation between standing in tests and in judgments is +.50, still this coefficient does not tell us how well the tests will pick out a *group* of the more capable students. For the coefficient +.50 is determined by relating the exact standings of each individual in the two series, standings in tests and standings in judgments.

Suppose the following standings of 16 individuals in the two series, the colon separating the number of individuals to be included in the 'good' group.

Standing in tests........	5	4	3	2	1	:	9	8	7	6	16	15	14	13	12	11	10
Standing in judgments...	1	2	3	4	5	:	6	7	8	9	10	11	12	13	14	15	16

Here, although the correlation by the method of rank differences is only +.61, the efficiency of the tests in picking out the 'good' group is 100 per cent. In other words, accuracy or inaccuracy of the internal arrangement of the groups is equally acceptable for the practical purpose of getting the best students into the advanced divisions. When tests are to be used in situations of this kind, they must be measured by how sharply they differentiate the 'good' as a group from the 'not-good' as a group. Such a measure would give the real practical value of the tests for the specific situation.

The measurement of a test by the true practical situation may be made by use of a formula published in 1907 by Karl Pearson.[1] Fortunately, the measure is exactly equivalent in meaning to the product-moment coefficient of correlation, and it is designated by the same symbol "*r*." Brown[2] in 1911 called attention to the formula. The exact description of the formula as it is stated in the title of the article in which it appeared is *A New Method of Determining the Correlation between a Measured Character 'A,' of which only the Percentage of Cases wherein 'B' Exceeds (or Falls Short of) a Given Intensity is Recorded for Each Grade of 'A.'* Thus the correlation between performance in tests and the general qualitative divisions of the group as judged may be determined; for the measured character '*A*' becomes the test measurements, and the character '*B*' given by alternative categories becomes the general qualitative divisions.[3] In the actual

[1] *Biometrika,* 7, 96–105.

[2] W. Brown, 'Mental Measurement.'

[3] Or vice versa, depending upon the information which is desired. If the alternative categories are based upon judgments, the coefficient will tell the value of the tests

situation, to be sure, there are usually three divisions—*good*, *mediocre*, and *poor;* but this difficulty is instantly overcome by dividing the group twice—once into the *good* and *not-good*, and again into the *poor* and *not-poor*. The formula is then applied for each division. In this way it is possible to tell whether the test works more efficiently in separating the *good*, or the *poor*, from the remainder of the group.

Important as the formula is in obtaining a measure of practical efficiency of a test, it has a still greater value. For it may be used to determine just where the division into classes should be made in order that the test may operate at its highest efficiency. Let us consider a group that has been divided into the *good* and *not-good* according to judgments. Clearly this division may be made at any point, and at each point of division there may be a different coefficient of correlation. If the changing values of the coefficient of correlation are plotted for the changing percentages of division (see charts) an excellent indication of the best point of division will be given. If the test is tried out for a sufficiently large number of individuals, 500 or more, this best point of division will be valid for all individuals of that class, and may be taken for use in all further work. The coefficient of correlation at this point of division is the indication of the true practical value of the test.

The presuppositions upon which the formula is based are linear regression—a presupposition for any correlation *coefficient*—and the Gaussian distribution in the alternative variable, here the judgments Strict normality is not essential for practical work but since the accuracy of the result will be affected by the distribution, the measure of the skewness of the curve should always be given whenever it can be computed.

The formula as stated by Pearson is

in selecting a certain percentage as judged. If the alternative categories are based upon tests, the coefficient will tell the relation between a certain percentage as tested and the judgments.

$$r = \frac{\dfrac{p}{\sigma_1}}{\dfrac{q}{\sigma_2}}; \qquad \frac{q}{\sigma_2} = \frac{\dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y/\sigma_2)^2}}{\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{y/\sigma_2}^{\infty} e^{-\frac{1}{2}y^2 dy}},$$

where p and q are the means of the two variates; and σ_1 and σ_2 are the two standard deviations.

In spite of the apparent complexity of the formula, the actual labor in application is very slight. The following ten simple steps will give the coefficient.

1. Determine the mean of the measured character.
2. Determine the mean of the members of the measured character included in the smaller class of the character given by alternative categories.
3. Determine the standard deviation of the measured character.
4. Subtract 1 from 2.
5. Divide 4 by 3.
6. Divide the number of cases in the smaller class by the total number of cases.
7. Subtract 6 from 1.00.
8. 7 equals $\frac{1}{2}(1 + a)$ of Sheppard's tables[1] of the probability integral; secure the corresponding z.
9. Divide 8 by 6.
10. Divide 5 by 9, which gives the correlation coefficient.

Tables of the probability integral (Sheppard's Tables) are essential.

The following problem illustrates the use that may be made of this formula.

Fifty college freshmen were given a series of mental tests. *Required*, (1) a measure of the efficiency of the combined tests; (2) the relative efficiency of the tests in separating the *good* and *poor* groups; (3) the percentage of individuals that should be included in each group in order that the tests may give the best results.

On the basis of the grades of the students, the group was

[1] *Biometrika*, II., or 'Tables for Statisticians and Biometricians,' edited by Karl Pearson.

divided into the *good* and *not-good*, by putting the highest eight per cent. in the *good* group. The correlation with standings in the combined tests was then computed by the formula described in this paper. Again the group was divided into the *good* and *not-good*, but this time the highest twelve per cent. were put into the *good* group. The correlation was found for this second division. Similar divisions were made at the 16, 20, 24, 28, 32, 36, 40, 44, and 48 per cent. points, and the correlation was computed at each division. A curve was then drawn showing the amount of the correlation at each of these points. (See chart.) A curve of the various degrees of correlation when the group is divided into the *poor* and *not-poor* was obtained in a similar way.

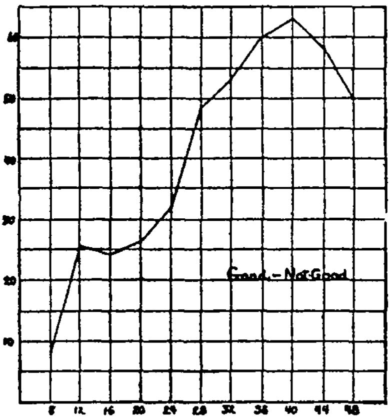

FIG. 1.

Vertical Axis—Coefficient of Correlation.

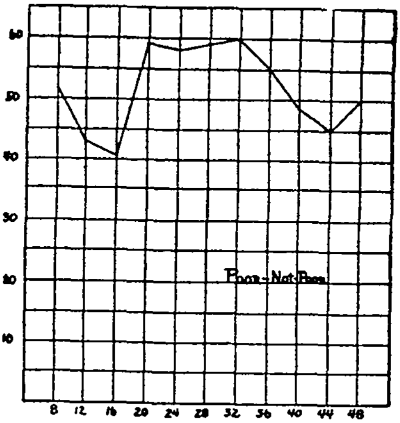

FIG. 2.

Horizontal Axis—Points of Revision.

From these two curves, it may be instantly concluded that if the *good* students are to be separated from the *not-good*, the separation had best be made at the forty per cent. mark, for here the correlation is the highest, +.63. In case the *poor* are to be selected, the division may be made anywhere from twenty to thirty-two per cent., and the correlation will be about +.58.

In contrast with these results, the only information which the product-moment formula gives from these data is that the correlation is +.43, *a value that is erroneous if the tests are to be used only for separating the individuals into homogeneous groups.*

This problem is one of a type that is common in education, business and industry. Wherever it is necessary to evaluate tests that are to be used in the selection of groups of individuals, the formula described in this paper should be used; for it is then possible to determine what the practical efficiency of the test really is, to weigh the relative accuracies of selecting the *good* or the *poor* individuals, and to determine the best percentage that can be included in either the *good* or *poor* groups.